|  |
| --- |
| **제11회「2023 빅콘테스트」결과보고서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 ☑ 표시 | | |
| **참가분야** | □ 생성형AI 분야 □ 데이터신기술 분야  ■ 정형데이터 분석 분야 □ 비정형데이터 분석 분야  □ 빅데이터플랫폼 활용 분야 | | |
| **세부리그**  \*해당시 체크 | ■ 어드밴스드 리그 □ 스타터 리그  \*정형데이터 분석분야에 한함(선택) | | |
| □ 지정주제 리그 □ 자유주제 리그  \*빅데이터플랫폼 활용분야에 한함(선택) | | |
| **개인/팀여부** | □ 개인 ■ 팀(총 4 명) | **개인/팀명** | Miraclassic de BOAZ |
| **지도교사명** | \*스타터 리그에 한함(선택) | | |
| **대표ID** | jwshin0908@gmail.com | | |

|  |  |
| --- | --- |
| **결과보고서 작성 안내 사항** | |
| **목차** | **Ⅰ. 분석 개요**  1. 주제 소개  2. 분석 배경 및 목적  3. 분석 순서도  **Ⅱ. 데이터 수집 및 전처리**  1. 데이터 수집  2. 데이터 전처리  3. 파생변수  **Ⅲ. 데이터 EDA**  1. 데이터 요약 및 시각화  2. 좌석 시각화  3. 가설 수립 및 검정  **Ⅳ. 콘서트홀 좌석 그룹핑**  1. 모델링  2. 좌석 그룹핑 결과  **Ⅴ. 공연 수요 예측 모델링**  1. 모델링 & 튜닝  2. 최적 모델 선정  3. 변수 중요도 해석  **Ⅵ. 가격 모델링**  1. 가격 모델링 개요  2. 가격 산출 공식  3. 가격 산출 예시  **Ⅶ. 결론**  1. 활용방안 & 기대효과  2. 의의 및 한계  **Ⅷ. 툴∙데이터 및 참고문헌**  1. 활용 툴∙데이터 및 참고문헌 출처 |

|  |  |  |
| --- | --- | --- |
| **Ⅰ.** |  | **분석 개요** |

**1. 주제 소개**

**가. 클래식 공연 활성화를 위한 가격 모델 수립**

클래식 공연 활성화를 위한 예술의전당 음악당 콘서트홀의 효과적 가격 모델을 수립한다. 본 분석은 크게 좌석 그룹핑과 공연별∙좌석별 가격 모델링으로 구성된다. 이때, 멤버십에 집중하여 가격 모델링을 진행한다. 가입한 예술의전당의 무료∙유료 회원을 기반으로 클래식 공연에 대한 관람객의 관심도를 수치화한다. 이후 관람객의 관심도에 따른 맞춤형 좌석과 가격을 제안한다. 이 과정에서 클러스터링을 통해 좌석을 그룹핑하고, 예측한 수요와 가격 산출 공식을 활용하여 각 좌석 등급에 맞는 가격을 도출한다.

**2. 분석 배경 및 목적**

**가. 분석 배경**

1. 공연시장에서의 클래식 공연

2023년 상반기 기준 전체 공연시장에서 클래식 장르 공연이 차지하는 비율은 38%로 KOPIS의 9가지 장르 중 가장 많은 공연이 이루어졌다. 그러나 티켓 판매 수와 티켓 판매액은 상대적으로 티켓 가격이 높은 뮤지컬 장르에서 가장 높았다. 클래식 공연의 티켓 판매 수는 13.5%, 티켓 판매액은 7.1%으로, 41.9%, 45%를 기록한 뮤지컬에 비해 비교적 낮은 편이다.

클래식 공연은 수요에 비해 공급이 많다는 지적이 나오고 있는 실상이다. 이러한 상황에서 관람객에게 합리적인 가격을 제공하여 클래식 공연의 수요를 높이고 현재의 많은 공급을 충족할 수 있어야 한다. 합리적인 가격 설정으로 공연에 대한 접근성을 높이고 예술의전당의 수익을 최대화하는 것이 필요하다.

1. 고객 관리의 중요성

고객과의 관계를 지속적으로 강화하기 위해서는 멤버십 가입 고객을 위한 맞춤형 관리가 필요하다. 고객 만족도 제고를 위해 멤버십을 고려한 가격 모델 수립 방안을 제안한다. 예술의전당은 현재 무료회원, 연령인증을 통한 싹틔우미∙노블 회원, 유료회원인 그린∙블루∙골드 제도를 시행 중이다. 회원제 사업 운영 이래 2022년에는 신규 유료 멤버십 증가율이 최대치를 기록하며 고객 관리의 필요성은 더욱 대두되었다. 신규 유료 멤버십의 증가는 기관 사업 활성화 및 예매율 상승에 기여한다. 관람객과 예술의전당의 지속적인 관계 유지의 중요성에 따라 멤버십에 초점을 맞추고자 한다.

**나. 분석 목적**

1. 다양한 고객을 수용하는 가격 모델

현재의 가격 모델은 다양한 관람객의 관심과 선호도를 충분히 고려하지 못한다. 이로 인해 일부 관람객들은 원하는 공연에 접근하는 데에 어려움을 겪는다. 이에 따라 관람객 특성과 공연의 성격을 반영한 새로운 좌석 그룹핑 기준을 수립하여 관람객들에게 더 나은 즐거움을 제공하고자 한다. 특히 가입한 멤버십에 따라 관람객의 관심도를 수치화하고 분석에 반영하여 다양한 관람객의 니즈와 관심도를 충족시킨다.

1. 합리적인 가격 산정을 통한 투명성 제고

예매과정에서 가격 산정에 대한 합리적 근거와 가격 정책의 투명성이 부족하다. 데이터 기반으로 합리적인 가격 모델을 구축하고 이를 보다 투명하게 공개하고자 한다. 결과적으로 관람객들에게 신뢰성 있는 정보를 제공하여 예술의전당 공연에 대한 신뢰 강화를 기대할 수 있다.

**3. 분석 순서도**

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

그림 1 분석 순서도

본 분석은 데이터 수집과 전처리, 장르별 좌석 그룹핑, 공연 수요 예측과 가격 제안 모델링 순서로   
진행된다.

|  |  |  |
| --- | --- | --- |
| **Ⅱ.** |  | **데이터 수집 및 전처리** |

**1. 데이터 수집**

**가. 제공 데이터**

1. 제공 데이터

제공 데이터에서 아래와 같은 컬럼을 활용하였다.

* 공연: 공연ID, 공연일자, 공연시간, 예매시작일, 선예매시작일, 러닝타임
* 예매: 공연ID, 좌석정보, 할인내역, 결제가격, 예매일자시간
* 대관: 공연ID, 공연일자, 공연시간, 장르
* 회원: 나이대, 성별, 멤버십, 회원여부

**나. 추가 데이터**

1. KOPIS API

KOPIS API를 활용하여 아래와 같은 데이터를 수집하였다. 제공데이터의 공연명이 가명정보처리가 되어 있어, 공연ID와 실제 공연을 매칭시키기 위해 공연 관련 데이터를 활용하였다. 추가로 해당 공연의 수요를 확인할 수 있는 티켓 판매 관련 데이터를 수집하였다.

* 공연: 공연명, 공연시작일, 공연종료일, 장르, 출연진, 제작진, 관람연령, 제작사, 티켓가격, 줄거리, 공연상태, 러닝타임
* 티켓판매: 공연일자, 티켓판매수

1. 예술의전당 공식 홈페이지

공연별 데이터 병합을 위해 예술의전당 공식 홈페이지의 공연일정을 기준으로 활용하였다.

* 공연일정: 공연명, 장르, 공연일자, 장소

**2. 데이터 전처리**

**가. 제공 데이터와 추가 데이터 매칭**

공연명을 인덱스로, 공연 세부 정보들을 컬럼으로 하는 데이터 구축을 목표로 두었다. 공연명과 공연일자 컬럼을 기준으로 데이터들을 매칭하기 위해 몇 가지 전처리 과정을 거쳤다. 먼저 공연시작일과 공연종료일에 속하는 기간만큼 행을 나누어 공연일자 단위 행으로 전처리하였고 컬럼명을 공연일자로 통일하였다. 전처리 이후 제공 데이터와 KOPIS API 두 데이터프레임에 공통으로 존재하는 공연명, 공연일자 컬럼을 key값으로 하여 두 데이터프레임을 병합하였다.

**나. 제공 예매 데이터**

1. 공연장소

콘서트홀에 한정하기 위해 리사이틀홀, IBK챔버홀 공연은 제거하였다.

1. 공연일자, 공연시작시간, 공연명코드

예매 데이터의 가명정보처리로 인해 데이터 매칭 오차가 발생하기 때문에 예술의전당 홈페이지 내 공연 세부 정보와 비교하여 실제로 공연하지 않은 데이터를 제거하였다. 이 과정에서 매칭되지 않는 공연 관련 컬럼은 제공 데이터의 컬럼 값을 사용하였다.

1. 멤버십종류 1~6

멤버십 종류별 컬럼을 생성하여 (0, 1) bool 형태로 표기하였다. 무료회원 중 연령 관련 회원인 싹틔우미, 노블의 기준에 맞지 않는 인적 사항을 가진 회원 데이터는 제거하였다.

1. 좌석번호

좌석 데이터를 분리, 층, 구역, 열, 좌석번호 컬럼을 생성하여 좌석에 대한 정보 구체적으로 표기하였다.

**3. 파생변수**

**가. 멤버십 관련 파생변수**

1. 관람객 관심도

관람객의 다양한 관심도를 가격 모델에 반영하기 위해 파생변수를 생성하였다. 멤버십에 따라 차등적으로 점수를 정의하여 클래식 공연에 대한 관람객의 관심도를 점수로 수치화한다. 비회원 0점을 기준으로 유료회원의 연회비를 점수 간격에 활용하였다. 회원가입을 한 일반회원은 1점, 연령인증을 거친 싹틔우미와 노블 회원은 2점으로 정의하였다. 유료회원인 그린, 블루, 골드회원은 각 연회비 2만원, 4만원, 10만원임을 고려하여 각 차액만큼을 더한 4점, 6점, 12점으로 수치화하였다. 예매 데이터의 각 멤버십종류에 따라 정의한 점수를 합계하였고 비회원 0점부터 모든 멤버십에 가입한 25점까지 분포한다.

1. 멤버십 등급

수치화된 관람객 관심도를 가입 멤버십 종류에 따라 S, A, B, C, D, E 6단계의 멤버십 등급으로 범주화하여 파생변수를 생성하였다. 이렇게 관람객 관심도를 등급으로 나누어 등급별로 선호 좌석, 선호 장르, 연령대 등 특징에 차이가 있는지 확인하고자 한다. 이후 해당 파생변수는 EDA와 좌석 선호도 시각화에 활용된다.

**나. 공연 관련 파생변수**

공연의 세부 정보를 면밀하게 분석하기 위해 관련 파생변수를 아래와 같이 생성하였다.

* 예매일자와 공연일자 간격: 공연일자보다 얼마나 빨리 예매했는지를 나타내는 파생변수
* 공연 월: 가정의 달 5월, 연말 12월 등 공연을 진행한 달을 나타내는 파생변수
* 선예매 여부: 선예매기간이 존재할 경우 1, 아닌 경우 0
* 좌석 좌표: 좌표평면 상에서의 좌석 위치에 대한 (x, y) 좌표
* 공연별 좌석 등급: 공연별 가격 종류 카운트한 후 좌석 등급 개수 집계

**다. 가격 산출 공식의 공연 관련 파생변수**

가격 산출 공식에 핵심적으로 활용한 공연 관련 파생변수를 아래와 같이 정의하였다.

1. 오픈 좌석 수

* 코로나 거리두기 정책에 따라 제한적으로 오픈된 좌석 수를 고려한 파생변수
* ~2020-06-07, 2021-07-01~ : 2,505석
* 2020-06-28 ~ 2021-06-30: 1,203 석

1. 티켓 판매율

* 공연당 티켓 판매율을 나타내는 파생변수
* 실제 판매된 좌석 수 / 오픈된 좌석 수

1. 멤버십 유입률

* 공연당 유료회원의 예매 정도를 나타내는 파생변수로 유입률이라고 명시
* 실제 판매된 좌석이 갖는 관람객 관심도의 총합 / (최대 관람객 관심도 \* 판매 좌석 수)

**4. 전처리 이후 최종 데이터**

1. 좌석별 데이터



그림 2 최종 좌석별 데이터

1. 공연별 데이터

텍스트, 영수증, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

그림 최종 공연별 데이터

1. 예매 데이터

텍스트, 영수증, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

그림 최종 예매 데이터

|  |  |  |
| --- | --- | --- |
| **Ⅲ.** |  | **데이터 EDA** |

**1. 데이터 요약 및 시각화**

**가. 전처리 이전 제공 데이터**

본격적인 분석 이전에 제공 데이터 EDA를 진행하였다. EDA 과정을 거치며 수많은 결측치들이 비회원 예매로부터 비롯되었음을 확인하였다.

|  |  |
| --- | --- |
| **EDA 시각화** | **설명 및 해석** |
|  | 나이  40, 50대의 예매 건수가 가장 많다. |
|  | 성별  여성이 남성의 약 2배 정도의 비율 차지한다. |
|  | 공연시작시간  주로 저녁 시간이며 19:30 이후의 공연이 큰 비율을 차지한다. |
|  | 러닝타임  90분에서 120분 정도의 러닝타임 공연 예매 건수가 가장 많다. |
|  | 장르  교향곡, 독주, 실내악 순서로 예매 건수가 많다. |

**나. 전처리 이후 예매 데이터**

|  |  |
| --- | --- |
| **EDA 시각화** | **설명 및 해석** |
|  | **멤버십 등급별 예매건수** |
|  | **연령대별 멤버십 등급 분포**  연령이 올라갈수록 높은 멤버십 등급을 보유한다. |
|  | **멤버십 등급별 선호 장르**  모든 멤버십 등급이 교향곡, 클래식을 가장 많이  예매한다. 상대적으로 높은 멤버십 등급은 독주를, 낮은 멤버십 등급은 합창과 성악, 오페라를 선호하는 경향을 보였다. |
|  | **멤버십 등급별 평균 할인율**  S등급이 평균적으로 가장 높은 할인율로 티켓을 예매하였다. 이어서 골드를 포함한 유료회원을  보유한 A, B등급의 평균 할인율이 높다. 비회원인  E등급의 대부분이 초대권 할인을 받아 높은 평균  할인율을 보인다. |
|  | **멤버십 등급별 평균 결제가격**  S, A등급이 정가가 높은 티켓과 좌석을 예매하지만 보유한 멤버십에 따라 큰 폭으로 할인을 받기  때문에 B, C등급보다 평균 결제 가격이 낮다.  비회원인 E등급의 대부분이 초대권 할인을 받아  평균 가격이 매우 낮음을 확인할 수 있다. |

**2. 좌석 시각화**

**가. 태블로를 활용한 좌석 배치도 구현**

|  |  |
| --- | --- |
|  |  |
| 그림 5 홈페이지 좌석 배치도(3층) | 그림 6 태블로 좌석 시각화 |

홈페이지 좌석 배치도 이미지의 1, 2, 3층을 구분하여 가상의 원점을 설정하고 각도 계산기와 선의 길이를 통해 원점에서부터 각 좌석의 각도와 간격을 측정했다.

원점과 좌석간 간격을 반지름으로 설정하고 각 각도에 맞게 부채꼴을 그리는 원리를 활용하여 부채꼴의 현 또는 호 위에 좌석 개수에 맞게 일정한 간격으로 점을 찍는 방식으로 좌석 좌표를 계산했다. 반복문을 활용하여 원점과 반지름 정보를 저장하고, 첫 번째 줄의 부채꼴 위 점을 찍고 반지름을 일정한 간격으로 증가시키며 그 다음 줄의 점을 찍어 x와 y좌표를 데이터프레임으로 저장하였다.

생성한 x, y 좌표 데이터를 활용하여 태블로로 좌석을 좌표평면에 마크를 원으로 설정하여 시각화하였다. 이후 판매량의 많고 적음이나 클러스터 구분 등 원하는 기준에 따라 색상에 차이를 두어 각 좌석의 특성을 시각화하는 데 활용하였다.

**나. 관람객 관심도에 따른 좌석 선호도**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 그림 7 S 등급 교향곡 예매 현황 | 그림 8 A 등급 교향곡 예매 현황 | 그림 9 E 등급 교향곡 예매 현황 |

장르 및 멤버십 등급별로 예매 데이터를 합하여 각 좌석별 총 예매 횟수를 통해 장르별로 각 멤버십 등급이 각 좌석을 얼마나 예매하였는지 확인하였다. 위 그림에서 밝은 색일수록 총 예매 횟수가 적은 좌석이고, 어두울 색일수록 총 예매 횟수가 많은 좌석이다.

교향곡의 예매 현황을 살펴보았을 때, S 등급 관람객은 약간 왼쪽으로 치우쳐진 앞쪽 좌석을 선호하고, A 등급 관람객 또한 왼쪽으로 치우쳐진 앞쪽 좌석을 선호하지만, 그 범위가 조금 더 넓게 나타난다. 비회원으로 구성된 E 등급의 경우 *특히* 선호하는 구역이 존재하기보다 전반적으로 모든 좌석에서 예매가 많았지만, 상대적으로 중앙과 뒤쪽 좌석 예매가 많다는 것을 확인할 수 있었다.

**다. 장르 및 코로나 여부에 따른 좌석 예매 현황**

|  |  |  |
| --- | --- | --- |
|  | **교향곡 예매 현황** | **합창 예매 현황** |
| **코로나 영향을**  **받았던 기간**  **(2020, 2021)** |  |  |
| **코로나 영향을**  **받지 않았던 기간**  **(2018, 2019, 2023)** |  |  |

예매 데이터 중 코로나의 영향을 고려해야 하는 2020년, 2021년의 데이터, 나머지 기간의 데이터를 구분, 총 예매 횟수를 나타내어 코로나 여부와 장르별로 좌석이 얼마나 예매되었는지 확인해보았다. 이전 예시와 마찬가지로 밝은 색일수록 총 예매 횟수가 적은 좌석이고, 어두울 색일수록 총 예매 횟수가 많은 좌석이다.

코로나 영향을 받은 기간은 좌석 띄어 앉기로 인해 좌석이 일정한 간격을 두고 예매되었다. 반대로 코로나 영향을 받지 않은 기간은 그러한 경향이 나타나지 않았으며, 모든 좌석 데이터를 바탕으로 장르별 선호된 좌석의 위치를 확인할 수 있었다.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 그림 9 교향곡 예매 현황 | 그림 10 클래식 예매 현황 | 그림 11 합창 예매 현황 |
|  |  |  |
| 그림 12 독주 예매 현황 | 그림 13 성악 예매 현황 | 그림 14 실내악 예매 현황 |

장르별로 전체 기간의 총 예매 횟수를 나타내 장르별로 선호하는 좌석에 차이가 있다는 것을 확인하였다. 예매 건수가 가장 많았던 장르 6개의 1층 좌석의 예시이다.

교향곡과 클래식은 약간 왼쪽에 치우쳐진 앞쪽 구역이 많이 예매되었고 B 블록의 오른쪽 끝 복도와 가까운 부분이 D 구역보다 선호되었다. 합창은 교향곡과 클래식보다 더 넓은 범위에서 중앙 부분의 좌석이 많이 예매되었다. 독주의 경우는 왼쪽에 치우쳐진 앞쪽 또는 중앙 부분의 좌석이 많이 예매되었고, 성악은 다른 장르에 비해 좁은 범위의 앞쪽 좌석이 선호되었다. 마지막으로 실내악은 다른 장르에 비해 좁은 범위의 앞쪽 중앙 부분의 좌석이 선호되었다는 것을 확인할 수 있었다.

**3. 가설 수립 및 검정**

공연과 관람객의 특성과 관련된 가설들을 검정 및 활용하였다.

1. 멤버십 가입에 따라

a. 공연에 대한 지불 용의가 다를 것이다.

b. 선호 장르가 다를 것이다.

c. 선호 좌석이 다를 것이다.

d. 이용하는 할인 내역이 다를 것이다.

* 가설 1.a와 1.b에 대해 독립성 검정을 시행해본 결과, p-value가 0에 가까운 값으로 나와 관람객 등급과 결제 가격, 선호 장르가 관련이 있다고 결론 내릴 수 있다.
* 가설 1.c와 1.d는 시각화를 통해 관람객 관심도에 따른 좌석과 할인내역이 뚜렷하게 다른 경향성을 보임을 확인하였다.

2. 장르에 따라 선호하는 좌석이 다를 것이다.

* 다양한 리뷰 및 컬럼, 선행연구에서 확인하였고 이에 따라 좌석 그룹핑을 장르별로 구분하여 진행하였다.

3. 일찍 예매된 좌석일수록 좋은 좌석일 것이다.

* 위의 사실을 가정하고 예매일자와 공연일자 사이의 간격을 좌석 그룹핑 모델에서 변수로 활용하여 분석하였다. 그 결과 좌석 등급에 유의한 영향을 미치는 것으로 파악하여 가설이 맞다고 검증하였다.

4. 선예매 공연은 인기가 많은 공연일 것이다.

* 직관적인 사실로, 선예매 여부를 수요 예측 모델 내의 독립변수로 활용하였고, 종속변수를 예측하는 데에 유의한 변수로 작용하였다. 이에 따라 맞는 가설임을 확인하였다.

|  |  |  |
| --- | --- | --- |
| **Ⅳ.** |  | **콘서트홀 좌석 그룹핑** |

**1. 모델링**

**가. 좌석별 데이터 생성**

1. 예매 데이터

예매 데이터를 개별 좌석을 기준으로 groupby하여 좌석별 데이터를 생성하였다. 예매 횟수는 합계, 예매 가격은 평균, 예매일자와 공연일자 간격은 평균, 멤버십 점수는 평균하여 groupby하였고, ‘총예매횟수’, ‘평균예매가격’, ‘평균예매일자공연일자간격’, ‘평균멤버십점수’ 컬럼을 생성하였다.

총 예매 횟수가 많은, 평균 예매 가격이 높은, 평균 예매일자 공연일자 간격이 긴, 평균 멤버십 점수가 높은 좌석이 ‘인기가 많은 좌석’이라고 정의하여 분석을 진행하였다.

1. X, Y 좌표 데이터

개별 좌석 데이터에 태블로에서 좌석을 구현할 때 활용한 X, Y 좌표를 추가하여 좌석의 위치도 고려하고자 하였다.

1. 최종 좌석별 데이터

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 15 좌석별 데이터

좌석별로 예매 데이터의 컬럼과 X, Y 좌표 데이터를 연결하여 위와 같이 인기도를 판단할 수 있는데이터와 위치 데이터가 결합된 최종 좌석별 데이터를 생성하였다. 모든 변수를 고려하고자 클러스터링을 진행하였다.

**나. 클러스터링**

장르별로 선호되는 좌석이 다를 것이라고 예상하여 좌석별 데이터를 장르별로 나누어 k-means 클러스터링을 진행하였다. 엘보우 그래프와 실루엣 계수의 평균값, 분포를 확인하여 최적 클러스터 개수 k를 선정하였다. 좌석별 데이터 중 예매 데이터에서 생성한 4개의 컬럼별로 순위값을 생성한 후 각 컬럼에서 모두 순위가 높은 클러스터가 인기가 많은 클러스터로 설정하였다.

텍스트, 그래프, 라인, 다채로움이(가) 표시된 사진

자동 생성된 설명

그림 16 교향곡 데이터의 엘보우 그래프와 실루엣 계수

교향곡 데이터의 경우, 엘보우 그래프와 실루엣 계수가 위와 같이 나왔기 때문에 k=5로 설정하였다. 같은 방식으로 클래식은 k=5, 합창은 k=4, 독주는 k=5로 설정하였다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 17 교향곡 클러스터의 인기도 컬럼별 순위와 순위평균

교향곡 클러스터링 결과를 보면, 순위평균을 기준으로 클러스터를 오름차순 정렬했을 때 4, 2, 0, 3, 1클러스터 순서로 순위가 높은 것을 확인할 수 있다. 따라서 4번 클러스터를 인기가 가장 많은 좌석 클러스터, 1번 클러스터를 인기가 가장 적은 좌석으로 판단하였다.

그 후 기존 좌석별 데이터에 클러스터 데이터를 연결하여 좌석 좌표에 클러스터를 나타내었다. 교향곡의 클러스터는 하단 사진과 같이 시각화되었다. 따뜻한 색상일수록 인기가 많은 좌석 클러스터, 차가운 색상일수록 인기가 적은 좌석 클러스터이다.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 그림 18 좌석 좌표에 나타낸 클러스터, 왼쪽부터 1층~3층 | | |

**2. 좌석 그룹핑 결과**

클러스터링을 통해 장르별 그룹핑한 결과는 다음과 같다.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 그림 19 교향곡 5개의 클러스터 | 그림 20 클래식 5개의 클러스터 | 그림 21 합창 4개의 클러스터 |

교향곡은 5개로 그룹핑되었고, 왼쪽으로 치우쳐진 앞쪽에 인기가 많은 클러스터가 존재했다. D블록보다 왼쪽에 위치한 B블록을 선호하는 것으로 나타났다. 협연자가 존재한다면 오케스트라를 등지고 지휘자의 왼쪽에 위치하기 때문에 이러한 경향이 두드러지는 것으로 보인다.

클래식의 경우 5개로 그룹핑되었으며 교향곡에 비해 더 넓은 범위의 중앙 부분에 인기가 많은 클러스터가 존재했고, 합창석이 인기가 많았다. 한쪽으로 치우쳐 앉으면 특정 악기의 소리만 강하게 들릴 수 있기 때문에 중앙 부분이 인기가 많은 것으로 보인다.

합창은 4개의 클러스터가 생성되었고 중앙 부분 넓은 구역을 선호하였다. 다양한 음성의 소리를 비슷한 거리에서 감상하는 데에 중앙 부분이 적합하기 때문으로 보인다.

|  |  |  |
| --- | --- | --- |
| **Ⅴ.** |  | **공연 수요 예측 모델링** |

**1. 모델링 & 튜닝**

**가. X, y 변수**

R(일반예매율), M(멤버십 유입률)을 예측하는 회귀모델 학습을 위해 공연별 데이터를 변형해 다음과 같이 일부 변수를 사용하였다.

* X : 7개 변수(공연 시작시간, 선예매여부, 장르, 내한여부, 러닝타임, 요일, 월)
* Y : R(일반예매율), M(멤버십 유입률)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

**나. 범주형 데이터**

범주형 데이터는 더미화(one-hot encoding)처리하였고, 3개 범주형 컬럼(장르, 요일, 월)으로 30개의 컬럼이 생성되었다.

**다. train/test 데이터 분리**

데이터셋에서 80%는 train, 20%는 test 데이터셋으로 분리하여 이후 Grid Search 진행 시 CV=5로 교차 검증(Cross Validation)을 진행하였다.

**라. 연속형 데이터**

연속형 데이터는 Standard Scaling을 거쳐 정규화하였다. train 데이터셋을 정규화한 만큼 test 데이터셋 정규화를 진행하였다.

**마. 모델 학습**

R(일반예매율), M(멤버십 유입률)을 예측하고자 각각의 예측 y별 회귀모델을 선택하였고, 아래의 모델들을 활용하여 학습을 진행하였다.

* Decision Tree: 데이터 분류와 예측을 위한 트리 구조의 알고리즘으로, 데이터의 특성에 따라 의사결정 규칙을 생성하여 문제를 해결
* Random Forest: 여러 개의 의사결정나무(Decision Tree)를 조합하여 안정적이고 정확한 예측을 수행하는 앙상블(Ensemble) 기법
* LightGBM: 일반적인 그라디언트 부스팅(Gradient Boosting) 방법을 사용하면서도 빠른 학습과 예측 속도, 과적합 방지 등의 성능 향상을 위한 최적화 기법을 적용한 부스팅 알고리즘
* XGBoost: 트리 분류기와 부스팅 알고리즘을 활용하여 높은 예측 성능을 가지며 과적합 방지와 빠른 학습 및 예측 속도 등이 특징인 알고리즘
* CatBoost: 범주형 특성을 다루는데 있어서 탁월한 예측 성능을 가지며 자체적으로 과적합을 방지하는 기능을 가진 부스팅 알고리즘
* MLP: 인공 신경망 기반의 다층 퍼셉트론으로 복잡한 비선형 문제를 다룰 수 있는 대표적인 머신러닝 알고리즘
* Ensemble: 여러 개의 머신러닝 모델을 결합하여 예측을 수행하는 앙상블 기법, 본 실험에서는 성능 상위 5개의 머신러닝 모델들을 결합함(Decision Tree 제외)

**바. 모델 하이퍼파라미터 튜닝**

이후 Grid Search CV를 진행하여 다음과 같이 모델별로 최적의 하이퍼파라미터를 선정하였다.

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **Random Forest** | **LightGBM** |
|  |  |  |
| **XGBoost** | **CatBoost** | **MLP** |
|  |  |  |

**2. 최적 모델 선정**

**가. R(일반예매율) 예측**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 모델  평가 지표 | Decision  Tree | Random Forest | LightGBM | XGBoost | **CatBoost** | MLP  Regressor | Ensemble  (상위 5개  모델 앙상블) |
| **MSE** | 0.0945 | 0.0730 | 0.0717 | 0.0726 | **0.0716** | 0.0721 | 0.0716 |
| **MSLE** | 0.0413 | 0.0322 | 0.0317 | 0.0323 | **0.0317** | 0.0324 | - |

위와 같이 모델을 비교해본 결과 CatBoost의 MSE가 가장 낮으므로, 일반예매율 예측 모델로 CatBoost를 선정하였다.

**나. M(멤버십 유입률) 예측**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 모델  평가 지표 | Decision  Tree | Random Forest | LightGBM | XGBoost | **CatBoost** | MLP  Regressor | Ensemble  (상위 5개  모델 앙상블) |
| **MSE** | 0.0157 | 0.0146 | 0.0146 | 0.0145 | **0.0144** | 0.0130 | 0.0140 |
| **MSLE** | 0.0117 | 0.0110 | 0.0109 | 0.0109 | **0.0985** | 0.0097 | 0.0105 |

위와 같이 모델을 비교해본 결과 MLP의 MSE가 가장 낮지만 모델의 해석력을 위해 멤버십 유입률 예측 모델로 CatBoost를 선정하였다.

**3. 변수 중요도 해석**

**가. R(일반예매율) 예측 : SHAP Feature Importance**

Shap (SHapley Additive exPlanations)란 ML 모델에서 각 변수를 하나씩 빼고 더하면서 해당 변수가 y 예측의 정확도에 얼마나 기여하는지 계산하여 각 변수의 중요도를 수치화하는 방법이다.

**나. R(일반예매율) 예측 : SHAP Summary Plot**

1. SHAP value 해석 방법

|  |  |
| --- | --- |
|  | 인과관계가 아닌 기여 방향의 경향성으로 해석해야 한다.  색상은 각 변수의 기여 방향을 나타낸다.  - 빨간색 : 변수 값이 높으면  - 파란색 : 변수 값이 낮으면  영역은 y 변수에 대한 작용 방향을 나타낸다.  - 양의 영역 : 예측 결과에 긍정적 요인  - 음의 영역 : 예측 결과에 부정적 요인 |
| 그림 22 SHAP value plot |

1. 변수별 SHAP value 해석 예시

* 요일\_금요일이면 일반예매율이 낮은 경향이 있다.
* 공연시작시간이 늦으면 일반예매율이 낮은 경향이 있다.
* 요일\_화요일이면 일반예매율이 높은 경향이 있다.
* 선예매여부이면 일반예매율이 높은 경향이 있다.
* 월\_6이면 일반예매율이 높은 경향이 있다.
* 요일\_수요일이면 일반예매율이 높은 경향이 있다.

**다. R(일반예매율) 예측 : SHAP Dependence Plot**

중요도 상위 15 변수 중 예측 기여 경향성을 살펴보자면 다음과 같다.

* 값이 커지면 일반예매율이 높은 경향이 있는 변수들: 요일\_화요일, 선예매여부, 월\_6, 요일\_수요일, 내한여부, 장르\_합창, 월\_3
* 값이 커지면 일반예매율이 낮은 경향이 있는 변수들: 요일\_금요일, 공연시작시간, 러닝타임, 장르\_독주, 월\_12, 월\_9, 요일\_목요일
* 해석 불가한 변수: 장르\_교향곡

**마. M(멤버십 유입률) 예측 : SHAP Summary Plot**

1. 변수별 SHAP value 해석 예시

|  |  |
| --- | --- |
|  | - 장르\_교향곡이면 멤버십 유입률이 높은 경향 있음  - 공연시작시간이 늦으면 멤버십 유입률이 높은 경향 있음  - 요일\_수요일이면 멤버십 유입률이 낮은 경향 있음  - 선예매여부이면 멤버십 유입률이 높은 경향 있음  - 장르\_독주이면 멤버십 유입률이 높은 경향 있음  - 요일\_금요일이면 멤버십 유입률이 높은 경향 있음 |
| 그림 23 SHAP value plot |

**바. M(멤버십 유입률) 예측 : SHAP Dependence Plot**

중요도 상위 15 변수 중 예측 기여 경향성을 살펴보자면 다음과 같다.

* 값이 커지면 일반예매율이 높은 경향이 있는 변수들: 장르\_교향곡, 공연시작시간, 선예매여부, 장르\_독주, 요일\_금요일, 월\_12, 요일\_토요일, 내한여부, 월\_3
* 값이 커지면 일반예매율이 낮은 경향이 있는 변수들: 요일\_수요일, 요일\_화요일, 장르\_합창, 러닝타임, 월\_6, 월\_2

|  |  |  |
| --- | --- | --- |
| **Ⅵ.** |  | **가격 모델링** |

**1. 가격 모델링 개요**

1. 공연 특성에 따른 클러스터링
2. 클러스터 내 평균 가격으로 기준 가격 설정
3. R(일반예매율)과 M(멤버십 유입률)로 기준 가격 보정
4. 최저가 - 최고가 범위 내에서 좌석 등급에 따라 가격 차등 분배

**2. 가격 산출 공식**

**가. 공연 특성에 따른 클러스터링**

1. 데이터

* 공연 정보: 공연시작시간, 선예매여부, 장르, 내한여부, 러닝타임, 요일, 월

1. 공연 클러스터링 및 결과 분석

공연에 대한 기준 가격을 설정하기 위해 공연 정보를 기반으로 K-Means 클러스터링을 진행하였다. 군집 개수에 따른 실루엣 계수 분석 결과, 군집이 6개일 때 가장 적절한 모습을 보였다.

다채로움, 스크린샷, 그래픽, 그래픽 디자인이(가) 표시된 사진

자동 생성된 설명

그림 24 공연 특성에 따른 클러스터링 실루엣 계수 분석

1. 6개 군집으로 클러스터링 및 군집 특성 분석

|  |  |
| --- | --- |
| 실루엣 계수 분석 결과에 따라 6개의 군집으로 클러스터링을 진행하였고, PCA를 활용하여 6개의 군집을 시각화하였다. |  |
| 그림 25 PCA를 통한 공연 군집 시각화 |

**나. 클러스터 내 평균 가격으로 기준 가격 설정**

공연이 할당된 클러스터 내의 최고가 평균과 최저가 평균을 각각 최고가, 최저가 기준 가격으로 선정하였다.

**다. 예매율과 멤버십 유입률로 기준 가격 보정**

|  |
| --- |
|  |
| 그림 26 기준 가격 보정 공식 (제안 가격 산출) |

1. (일반 예매율), (멤버십 가중 예매율), (멤버십 유입률) 산출 공식

* =일반 예매율 = 판매 좌석수 / 오픈 좌석수
* =멤버십 가중 예매율 = 전체 관람객 관심도 합계 / (25 오픈 좌석 수)
* =멤버십 유입률 = 전체 관람객 관심도 합계 / (25 오픈 좌석 수)

1. 일반 예매율 가중 함수, 멤버십 유입률 가중 함수

공연 자체에 대한 수요를 의미하는 일반예매율과 유료회원의 유입 정도를 의미하는 멤버십 유입률을 예측하여 가격에 반영할 수 있도록 가중 함수를 제시한다. 공연 클러스터의 평균 가격, 즉 기준 가격에 가중함수를 곱하여 새로운 가격을 제안한다. 이때 일반예매율과 멤버십 유입률이 평균보다 높을수록 가격을 인상하도록, 낮을수록 가격을 인하하도록 설정하였다.

|  |
| --- |
|  |
| 그림 27 가중 함수 공식 |

|  |  |
| --- | --- |
| **(일반 예매율) 가중 함수** | **(멤버십 유입률) 가중 함수** |
|  |  |
|  |  |

텍스트, 폰트, 스크린샷, 친필이(가) 표시된 사진

자동 생성된 설명

그림 28 제안 가격 산출 공식

**라. [최저가 – 최고가] 범위 내에서 좌석 등급에 따라 가격 차등 분배**

1. 좌석 비율에 따라 가격 차등 분배

각 장르에 따라 분류된 좌석 등급에 제안 가격을 차별화하여 분배하고자 한다. 각 좌석 등급에 대한 가격 차별화에 좌석 수 차이를 활용하여 아래와 같은 과정으로 진행한다.

1. (최고가 - 최저가) 가격 차이 산출
2. 등급별 좌석 개수 차이 집계
3. 좌석 차이 비율 = 좌석 개수 차이 / 좌석 개수 차이 합산
4. 좌석 차이 비율에 따라 가격 차이 차등 분배하여 등급별 가격 최종 선정
5. 가격 차등 분배 방식 예시 및 도식화

5개의 좌석등급으로 분류된 아래와 같은 예시가 있다고 가정하자. 이때 기준 최고가인 10만원, 기준 최저가인 1만원과 등급별 좌석 수를 활용하여 아래와 같이 최종적인 좌석 등급별 가격을 제안한다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 29 가격 차등 분배 방식 예시 및 도식화

**3. 가격 산출 예시**

**가. 예시 : 2023.10.19 클래식 공연**

1. 가격을 산출하고자 하는 공연 정보 수집

분석에 활용한 과거 데이터가 아닌 그 이후에 진행될 공연 데이터를 예시로 모델에 적용해보았다.

|  |  |
| --- | --- |
| **변수** | **값** |
| 공연일자 | 2023-10-19(목) |
| 공연시작시간 | 1930 |
| 선예매여부 | 1 |
| 예매시작일 | 20230823 |
| 장르 | 클래식 |
| 휴게시간 | 0 |
| 공연명 | 예술의전당 전관 개관 30주년 특별음악회-코리안챔버오케스트라 초청 콘서트 |
| 내한여부 | 0 |
| 러닝타임 | 2시간 |

1. 예측 𝑅(일반예매율), 𝑀(멤버십 유입률) & 할당 클러스터 기준가 예측

공연 수요 예측 모델을 활용하여 일반예매율과 멤버십 유입률을 예측한다. 공연 클러스터에 할당하여 기준 최고가와 최저가를 산출한다.

|  |  |  |  |
| --- | --- | --- | --- |
| **R(일반예매율)** | **M(멤버십유입률)** | **cluster\_R석** | **cluster\_A석** |
| 0.51146 | 0.12985 | 81696.96970 | 25454.54545 |

1. 최고가 ∙ 최저가 계산 & 장르별 좌석 등급별 가격 차등 분배

가격 산출 공식에 대입하여 최고가 제안가격과 최저가 제안가격을 도출한다. 클래식 장르의 좌석 그룹핑은 5단계이고 각 좌석 등급별 좌석 수는 아래와 같다. 도출된 최고가 제안가격과 최저가 제안가격을 좌석 등급별로 차등 분배한 결과 최종 가격은 하단의 오른쪽 표와 같이 제안할 수 있다. 나아가 해당 제안가격을 천 원 단위로 보정하여 활용할 수 있다.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **변수** | **값** | | 제안가격\_최고가 | 80940.83014 | | 제안가격\_최저가 | 25218.95301 |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **클래식 좌석** | **R석** | **A석** | **B석** | **C석** | **D석** | | 개수 | 194 | 234 | 490 | 674 | 915 | | |  |  |  | | --- | --- | --- | | **좌석 등급** | **가격(원)** | **개수(개)** | | R석 | 80,941 | 194 | | A석 | 77,850 | 234 | | B석 | 58,065 | 490 | | C석 | 43,845 | 674 | | D석 | 25,219 | 915 | |
| 그림 30 제안 최고가 및 최저가, 등급별 좌석수 | 그림 31 최종 등급별 제안 가격 |

|  |  |  |
| --- | --- | --- |
| **Ⅶ.** |  | **결론** |

**1. 활용방안 & 기대효과**

**가. 활용방안**

1. 가격 산출 모델 자동화

공연의 세부 정보 데이터프레임을 모델에 넣어 좌석 그룹핑과 각 좌석 등급별 가격 산출을 자동화할 수 있다.

1. 좌석 예매 통계 제공 서비스

|  |  |
| --- | --- |
| 그림 32 예매 통계 제공 서비스 예시 | 좌측의 예시 사진과 같이 좌석을 선택했을 때 좌석 등급과 함께 ‘어떤 고객이 얼마나 예매했었는지’와 같이 좌석에 대한 추가적인 데이터를 제공하는 데에 활용할 수 있다. 이를 통해 가격이 설정된 근거를 제공하여 신뢰성을 확보하고, 좌석을 선택하는 데 활용할 수 있는 추가적인 데이터를 제공하여 관람객의 구매 결정에 도움을 줄 수 있을 것이다.  예를 들면 S~E등급의 고객이 해당 좌석을 얼마나 예매했었는가를 보여준다면 해당 좌석의 등급이 설정된 근거를 제시할 수 있으며, 관람객이 갖고 있는 공연에 대한 관심도나 지불용의에 맞게 좌석을 선택할 수 있을 것이다. |

**나. 주체별 기대효과**

1. 예술의 전당

예측 수요에 따라 가격을 보정함으로써 수요와 수익을 최대화하는 최적의 가격을 선정할 수 있다. 또한 합리적인 가격 및 좌석 구분을 통해 고객 만족도를 향상시키고, 공연장 재방문을 유도, 결론적으로 및 클래식 공연을 활성화할 수 있을 것이다.

1. 관람객

관람객들은 다양한 관심도를 가지고 있는데 클래식 공연에 대한 관심도에 따라 적절한 좌석을 선택할 수 있는 가이드라인을 제시할 수 있다. 합리적인 가격 체계를 마련하여 개개인의 지불 용의 내에서 최대의 소비 효용을 달성할 수 있다.

**2. 의의 및 한계**

**가. 의의**

일반예매율과 더불어 멤버십을 고려한 유입률 등 다양한 파생변수를 가격 모델에 활용하여 다양한 관람객의 관심도와 지불용의를 고려하고자 하였다. 태블로에서의 좌석 시각화를 통해 직관적으로 멤버십 등 다양한 요소에 따른 좌석의 위치별 선호도를 파악할 수 있다. 또한 멤버십을 활용하여 공연에 대한 관심도가 다양한 관람객에 맞춤형으로 대응할 수 있는 가격을 제안하였다.

**나. 한계**

공연의 인기도에 영향을 미칠 수 있는 출연진과 기획사의 인지도를 파악할 수 있는 데이터가 부재하였다. 하지만 이후에 출연진과 기획사의 인지도를 포함하여 공연의 특성을 더 잘 반영한 가격 산출을 통해 모델을 고도화할 수 있다.

원본 데이터의 가명정보처리로 인해 공연 데이터의 매칭에 오차 발생하였다. 그러나 본 분석의 프레임워크를 원본 데이터셋에 적용하면 합리적인 가격이 산출될 것으로 기대된다.

|  |  |  |
| --- | --- | --- |
| **Ⅷ.** |  | **툴∙데이터 및 참고문헌** |

**1. 활용 툴∙데이터 및 참고문헌 출처**

**가. 툴**

1. Tableau

* 데이터 시각화

1. Python

* 데이터 시각화, 분석, 모델링
* pandas, numpy, matplotlib, seaborn, scipy, scikit-learn, seaborn, skimage, shap 등

**나. 데이터**

1. KOPIS API

* 제공 플랫폼 : KOPIS, 공연예술통합전산망
* 데이터 상세 : 공연목록, 공연상세, 일별 티켓판매수 및 티켓판매액

1. 예술의전당 공연일정

* 제공 플랫폼 : 예술의전당
* 데이터 상세 : 장르, 공연명, 기간, 장소

**다. 기사/칼럼**

* "30만~40만원으로 껑충" 비싸진 클래식 공연티켓값, 원인은 '항공 운임'
  + <https://www.asiae.co.kr/article/2022070422072908985>
* 日보다 70% 비싼 韓클래식 티켓값…"문화강국 되려면 문턱 낮춰야"
  + <https://www.hankyung.com/life/article/2022070610091>
* 음악회의 좋은 자리를 찾는 꿀팁!
  + <https://brunch.co.kr/@truth-art/24>
* 클래식 음악회 공연장별 꿀좌석은 어디?
  + <https://www.khan.co.kr/culture/culture-general/article/201908301402011>
* [칼럼] 국내 주요 클래식 콘서트홀 리뷰1 - 예술의전당 콘서트홀
  + <https://www.artinsight.co.kr/news/view.php?no=65604>

**라. 논문**

* 오경지. (2013). <클래식 음악공연의 좌석등급 및 티켓 가격 인플레 문제 연구>. 이화여자대학교공연예술대학원 석사학위논문.
* Ye, Peng, et al. "Customized Regression Model for Airbnb Dynamic Pricing." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, 2018, pp. 932–940.

**마. 보고서 등 기타 자료**

* Airbnb Dynamic Pricing Optimization
  + <https://github.com/tule2236/Airbnb-Dynamic-Pricing-Optimization/tree/master>
* 7 Common Pricing Models
  + <https://www.indeed.com/career-advice/career-development/pricing-modeling>
* Dynamic Pricing using ML
  + <https://medium.com/@baabak/dynamic-pricing-using-machine-learning-5e882282effe>
* Implementing Dynamic pricing strategy in python
  + <https://induraj2020.medium.com/implementing-dynamic-pricing-strategy-in-python-part-1-5bd7e1a1f382>
* Ticket Sales by Seat Type Report for Sports Organizations and Venues
  + <https://www.solverglobal.com/report-budget-forecast-and-dashboard-template-glossary/ticket-sales-by-seat-type-report-for-sports-organizations-and-venues/>
* Analysing Cinema Seating Patterns
  + <https://medium.com/analytics-vidhya/popcorn-data-analysing-cinema-seating-patterns-part-ii-987fbde9d363>
* 하우스 매니저가 알려주는 음악당 객석 꿀팁
  + <https://www.youtube.com/watch?v=k0kALhTkbKo>